新型人机关系下的人机双向信任

解煜彬 [1,2] 周荣刚 [1,3,4]

([1] 北京航空航天大学经济管理学院, 北京 100191) ([2] 香港城市大学工学院系统工程系, 香港 999077)

([3] 数据智能与智慧管理工信部重点实验室，北京 100191) ([4] 低碳治理与政策智能实验室，北京 100191)

摘要 随着人工智能技术的迅猛发展，人类与机器的互动频率不断增加，互动方式也日益复杂。传统的人机信任模型多聚焦于单向信任，即人类对机器的信任。然而，随着智能系统逐渐具备自主性和决策能力，人机信任的双向性逐渐成为研究的核心议题。本研究在回顾近年来人机信任理论模型的基础上，提出了基于倾向信任、感知信任和行为信任的人机双向信任理论结构模型，特别强调了“感知信任”作为人机互信的交互渠道的重要作用。此外，本文系统梳理了人机信任测量与计算建模方法的最新进展，重点探讨了机器对人类信任的测量方法及其实践意义，并提出了未来研究方向，以期为人机协作领域的理论发展和技术应用提供新的视角和指导框架。

关键词 人工智能，人机互信，信任，信任测量，人机团队

分类号 B849:C93

1. 引言

有效团队合作离不开信任(Alhaji et al., 2024; Pitardi & Marriott, 2021)。人类与智能机器之间的相互信任同样被视为成功合作的基础(Kobayashi et al., 2016; Lyons et al., 2019)。已有研究指出，不当或不足的信任会对团队合作产生负面影响(Walliser et al., 2019; Zhou et al., 2021)。例如，不恰当的信任可能导致人机关系恶化并损害团队效率(Olson & Xu, 2021)。随着 AI 和智能技术在社会生产和生活中的广泛应用，研究者们逐渐认识到，人工智能在人机协作中的角色[1]正不断增强，人机关系也正从传统的“辅助从属”向“平等合作”甚至“融合共生”的方向发展(Inga et al., 2023; Mueller et al., 2020)。在这一转变过程中，人机关系呈现出从单向性向双向性发展的趋势(Shi et al., 2024; Walliser et al., 2019)。由于人机关系具有双向性，因此培养人类与 AI 之间的相互信任是至关重要的。过去十年中，学界对人类对 AI/机器的信任展开了大量研究(de Visser et al., 2018; Jian, 2000; Kaur et al., 2020; Lee & See, 2004)，但对 AI/机器对人类的信任探讨相对有限。

人工智能对人类的信任基于共享心智模型(Shared Mental Model)的假设(Cuzzolin et al., 2020; Yuan et al., 2022)。美国国防高级研究计划局于 2015 年启动的可解释人工智能(XAI)项目提出人工智能心理解释模型，强调 AI 心智与人类心智一致性的重要性(Gunning et al., 2021)。Murphy (2024)同样指出，探究机器人如何智能地推断人类的信念、愿望和目标并采

---

[1]收稿日期：2024-10-12

* 航空科学基金(2024Z074051003)、国家自然科学基金(72171015, 72021001)和未来区块链与隐私计算高精尖创新中心资助。

通讯作者：周荣刚，E-mail: zhrg@buaa.edu.cn

取适当行动，是塑造可解释 AI 的关键。机器人需建立心智模型，才能与人类有效沟通与合作，从而获得人类的信任。为此，机器人不仅应理解人类的信任模式，还须以符合人类心理感知的方式传递信任，这被视为 AI 与人类实现心理协调的关键。基于此可见，信任这一心理变量在人机对齐中扮演着重要角色。齐玥等人(2024)率先关注了在人工智能时代人机信任关系的转变，强调 AI 对人类信任的重要性。他们在研究中借鉴人际信任模型及人-AI 信任理论框架，提出 AI 可通过主动感知自身与用户状态来评估对人类的信任水平，从而决定控制权的归属。这一研究首次对 AI 对人的信任进行了系统性定义，为深入探讨人机信任机制奠定了理论基础。然而，该文对于人类对 AI 的信任感知和双向信任的传递与测量的讨论仍显不足。

探讨机器对人的信任，尤其是人类对机器信任的感知，事关人机交互与技术应用的重要问题。尽管尚无法确定 AI 是否真正“信任”人类，但 AI 可以通过评估个体的可信度，并根据不同的可信度水平采取相应的行为，从而传递信任的信号。随着技术的不断进步，机器在某些任务中的能力甚至已经超越了初级员工(Babashahi et al., 2024)。在人机平等合作的新范式下，当机器识别出人类能力不足时，是否应主动提供建议或介入操作，以及以何种方式与深度介入，成为亟待解决的交互设计难题。例如，汽车的主动安全与主动避让技术往往在系统触发时强制接管人类操作。尽管此举可在某些场景下显著提升效率与安全性，但机器人设计三原则之一强调“机器人必须服从人类命令”，保留人类在伦理与法律层面上的最终决策权，包括对 AI 系统的“开关”和“授权”控制(Murphy & Wood, 2009; Jarrahi, 2018)。人类对 AI 的信任决定了其使用意愿，而这种信任通常与系统的性能表现紧密相关(Basu & Singhal, 2016)。反之，AI 对人类的信任决定是否将某些任务交由人类完成，取决于人类能力和表现能否满足系统预期。在人机交互中，人类对 AI 信任的主观感知影响是否行使“开关”等最终控制权。类似人类团队，当一方感受不到对方信任时便会增加不确定性与不安全感，从而引发防备心理乃至减少互动(Kramer, 1999)。同理，当人类感受不到 AI 的信任时，往往会对系统的可靠性和合作意图产生质疑，降低对系统的依赖与接受度。由此可见，双向信任的缺失不仅削弱合作关系，也可能阻碍技术的普及与应用。人类对 AI 的信任感知，不仅是技术设计的核心议题，更是推动 AI 长期发展的关键要素。

基于上述认知，我们的文献研究首先聚焦机器信任行为如何向人类传递“被信任”的感知，认为这一信任感知构成了人机双向信任的互动通道，并计划提出一个基于人类倾向、感知与行为的人机双向信任理论模型。其次，信任测量是当前人机信任研究的核心议题。尽管已有文献在一定程度上关注了机器信任，但尚未形成清晰的创建与度量框架，缺乏系统梳理与总结。为此，我们的第二个重点聚焦人机互信的测量方法，尤其是机器信任的倾向、感知、行为及综合计算模型等方法的深入探索。以上各方面将共同构筑我们关于人机双向信任的完整研究框架。为系统解答上述问题，我们首先收集了与人机信任相关的研究论文，具体收集与筛选要求见表 1。随后，本文将依次讨论：(1)人机信任的演变历程与当前挑战；(2)双向信

任模型及其理论基础；(3)人机双向信任的测量与建模方法；(4)相关应用案例分析；(5)未来研究方向。我们将在此基础上提出一个综合考虑人类信任心理结构、人机关系、人机交互行为及双向信任传递机制的人机互信理论与应用模型，并展望未来研究。

表 1 检索流程与规则

要点	详情
最后检索时间	1952~2024 年
数据来源	(1) Web of Science 核心集合，SCI & SSCI 数据库 (2) 中国知网数据库
关键词范围	“信任(Trust)”、“人际信任(Interpersonal Trust)”、“团队信任(Trust in Team)”、“人机信任(Human-Machine Trust)”、“人机互信(Human-Machine Mutual Trust)”、“人工智能信任(AI Trust)”、“机器人信任(Robot Trust)”、“算法信任(Trust in Algorithm)”、“信任倾向(Propensity to Trust)”、“自动化信任(Trust in automation 或 Automation Trust)”、“信任建模(Trust Modelling)”
筛选依据	(1) 将研究方向限定为心理学、管理学、社会学、计算机科学的相关方向。 (2) 选用英文或中文撰写的研究论文、会议论文和评论性文章，排除了书籍及书籍章节。 (3) 限定学科领域，删除无关与重复文献。 (4) 进一步筛查人机信任研究相关的重点期刊后，逐个阅读文章标题、摘要，对引用的重要研究进行整理和筛选。
筛选结果	134 篇相关文献

2. 人机信任的演变历程与当前面临的挑战

2.1 人机信任的演变历程

信任是一个多维度的概念，在心理学、社会学、经济学以及人机交互等多个领域都有广泛应用(Lee & See, 2004; Mayer et al., 1995; Simpson, 2007)。最初，人机信任主要关注人类对机器的信任：Muir (1987)探讨了在自动化系统中人类对机器信任的建立与丧失过程；Lee 和 See(2004)提出自动化信任的定义，将其视为一种在不确定情境下与代理人协作的态度，并引发了对情感、态度及情境等因素对信任影响的关注(Waytz et al., 2014; Schaefer et al., 2014; Hoff & Bashir, 2015；Merritt et al., 2013)。这一阶段，人机信任理论逐渐得到了众多学者的实证验证(Khastigir et al., 2017)。此后，研究进一步聚焦自动驾驶汽车(Choi & Ji, 2015; Robertson, 2021; Zhang et al., 2019)、人工智能代理(Regli & Annighoefer, 2022; Wang & Moulden, 2021)和航天器(Kintz et al., 2023)等领域的应用研究中，以及在动态环境下的人机信任校准和修复(Schaefer et al., 2016; Kaplan et al., 2023)。

随着机器具备自主学习与交互能力，信任从单向依赖走向双向互信(Chung et al., 2024)。de Visser 等人(2018)首次强调机器对人类的信任的重要性；Azevedo-Sa 等人(2021)提出了“人工信任” (Artificial Trust)的概念，构建了兼顾人对机器信任和机器对人信任的人机双向信任模型；Jorge 等人(2024, 2022a, 2022b)进一步拓展了 AI 信任的内涵，指出其不仅包含对人类能力与意愿的评估，还与感知成本、收益等因素相关；同时，他们定义了机器对人类信任的

结构并提出测量方法。齐玥等人(2024)明确了 AI 对人信任的关键影响因素，并构建了人-AI 互信模型。整体来看，人机信任正由静态单向向动态双向演进；随着机器智能化水平提高，对个体心理与认知差异的研究也愈发重要。然而，人机互信关系的系统探讨、双向信任传递机制等方面仍需进一步研究。

2.2 人工智能时代的新型人机关系

传统的工具论将机器视为从属工具，但随着人工智能日益增强的自主性与复杂性(de Visser et al., 2018; Yang et al., 2023)，人机关系正由“工具型”向“合作伙伴型”甚至“共生型”转变 (许为，葛列众，2020)。新型人机关系主要体现为：1. 从辅助到协作：传统人机关系以人工智能辅助人类完成任务为主，但如今，AI 已具备自主感知、决策和学习的能力，能够参与复杂任务的分工与协作。2. 人机团队结构：Sycara 和 Lewis(2004)提出了机器在团队中的三种角色：支持个人任务、作为平等成员或辅助整个团队。在新型人机关系中，AI 也被视为人类团队的代理替代品，可充当“合成型人类”(McNeese et al., 2018)。3.双向交互与信任：在新型人机关系中，如何让用户接受和采用人工智能系统，以及人工智能如何被视为队友而非工具，成为其融入人类团队的核心挑战(Wong et al., 2024)。信任被认为是影响上述问题的核心心理变量(Caldwell et al., 2022; Georganta & Ulfert, 2024; Nies, 2009; Ulfert et al., 2024)。这一新型人机关系的核心特征是双向交互，这要求研究者进一步聚焦人机双向信任的定义、测量方法及其动态变化。

2.3 新型人机关系下人机信任的挑战

在新型人机关系中构建信任面临多重挑战，尤其在高度依赖人工智能的协作场景下，双向信任的定义与测量尤为关键。尽管已有研究从人际信任视角探讨了人机信任的核心内涵(许为等, 2024; 齐玥等, 2024; Jorge et al., 2022a, 2022b)，但仍存在以下研究空白：

1. 双向信任的交互与传递机制缺乏心理层面的依据。目前研究主要聚焦定义与结构模型，缺乏对双向信任心理机制的探讨，尤其是信任在情感、信息与决策交流中的传递过程。例如，人类对 AI 信任的感知及其信任需求。本文将结合人际与人机信任模型，提出基于人类感知“被信任”的双向信任模型，深入分析心理传递机制。

2. 缺乏对个体特征差异的理解。在人类对机器的信任中，情绪、态度和性格的差异显著影响信任的效果。个体的信任倾向和算法厌恶等特质，可能成为信任建立的潜在障碍。本文拟将这些特质纳入分析，重点探讨它们对信任互动与传递的作用。

3. 人机双向信任的测量方法缺乏系统的梳理，尤其缺少对机器信任的测量方法。现有方法多聚焦于问卷调查、心理感知和行为观察，对人类信任机器的测量较成熟，但对机器信任人类的测量维度与方法尚未完善。本文将整合现有方法，构建适用于双向信任的测量体系。

4. 对双向信任的影响因素与作用效果的研究不足。当前研究对双向信任的影响因素与其作用效果的探索，尤其在实验和实证层面仍显匮乏。本文将重点分析双向信任的关键变量及其作用机制，旨在弥补现有研究的不足，完善相关理论与实践框架。

3. 基于信任感知的双向信任模型

我们以信任理论模型的发展过程为主线，在系统梳理信任不同发展阶段的基础上，结合人际团队关系中的信任理论和人机互信理论模型，提出了一种基于信任感知的双向信任模型。该模型旨在阐明人机双向信任的发展过程及其相关影响机制，为构建更加高效、可靠的人机协作体系提供理论支持和实践指导。

3.1 人机互信的框架要素

信任理论的早期研究由 Mayer 等人(1995)提出，基于人际信任的结构，构建了涵盖能力(competence)、善意(benevolence)和正直(integrity)的三因素模型。随后，Lee 和 See(2004)从自动化信任的角度提出了信任的三大基础：表现(performance)、过程(process)和目的(purpose)。Earle 和 Siegrist(2006)则将信任划分为关系信任与计算信任，强调了关系的重要性。Merritt 和 Ilgen(2008)引入了倾向性信任的概念，提出信任是介于倾向性信任(dispositional trust)和历史信任(history-based trust)之间的连续体。随后，Hoff 和 Bashir(2015)构建了包含倾向性信任、情境信任(situational trust)和习得信任(learned trust)的三层人机信任模型。高在峰等人(2021)以自动驾驶汽车为研究对象，提出了基于信任发展过程的动态信任框架，明确了信任的四个层面：倾向性信任、初始信任、实时信任与事后信任。近年来，随着人工智能的自主性与复杂性提升，研究者开始关注机器对人的信任。Jorge 等人(2022a)基于 Mayer 等人(1995)的框架，将人类可信度划分为三个维度：能力(Competence)——个体完成任务的成功程度；善意(Benevolence)——个体愿意无私帮助其他智能体，而非损害其目标的意愿。正直(Integrity)——个体是否展现出真实、诚信和符合道德原则的行为。他们的理论基础主要聚焦于构建可信赖的人类的评价体系。齐玥等人(2024)同样充分考虑人际信任模型的结构，基于信任的过程和状态，提出了包含初始阶段、感知阶段和行为阶段的人与 AI 动态互信模型。除了初始信任和行为信任，他们还特别强调了感知阶段的重要性，并将其进一步细分为对系统状态的感知和对用户状态的感知。

现有研究通常将信任划分为两层：倾向性信任是一种固有的信任特质，与具体情境无关，具有高度稳定性；历史信任则通过交互过程生成，受情境影响并动态调整。研究者普遍认为信任是一个动态发展的过程，涵盖初始信任、感知信任和事后信任。纵观现有模型，研究者多将人机信任视为可信度的概念，专注于构建可信的人类与可信的 AI 评价体系。尽管可信行为通常是可观察的，但个体行为也受到对“感知被信任”的影响 (Salamon & Robinson, 2008)。然而，目前的人机信任框架，尤其在人机互信的交互感知方面，鲜有系统性的理论模型。信任作为一种心理特征，其在倾向信任与行为信任之间的情感传递性尚缺乏明确的模型支持。例如，人类对 AI 信任行为的感知这一关键维度，仍未得到充分关注。在人类团队中，已有研究证明信任的运作仅在人类能够感知到他方信任行为时发生(Baer et al., 2015; Hieronymi, 2008)，但关于人类对 AI 行为信任感知的研究仍较为匮乏。

基于这一视角，本研究提出了一个动态演化的三阶段人机互信模型，分为倾向信任、感知信任和行为信任三大核心阶段（如图 1 所示）。模型强调感知信任作为倾向信任与行为信任之间的关键桥梁，突出其在 AI、智能体与人类之间的传递作用。倾向信任：信任的初始阶段，源于个体固有特质，与具体情境无关，为后续信任发展奠定基础。感知信任：在交互过程中逐步形成，体现对另一方行为、态度和信任的动态感知，是信任情感传递与动态调整的核心。行为信任：信任的最终体现，通过具体的依赖、合作和行动表现，是基于行为反馈的事后信任，反映信任关系的最终结果。本模型突出了信任在交互中的动态演化过程，从倾向信任到行为信任的逐步发展，并深入揭示了信任在双向交互中的传递机制，为构建高效的人机协作关系提供了新理论视角和实践依据。

本模型的优势主要体现在以下几个方面：1.动态演化特性，即模型全面展现信任从倾向信任到感知信任再到行为信任的动态发展过程，适应人机交互中信任关系的复杂性与变化性；2.双向信任传递，即模型重点关注人类与智能体之间的双向互动，系统性强调感知信任在倾向信任与行为信任之间的桥梁作用，深入解析其在信任情感传递与动态调整中的关键意义，为人机交互优化提供独特指导；3.倾向信任的扩展视角，即在倾向信任阶段，引入算法信任视角，探讨算法初始信任来源及个人算法厌恶倾向的影响，为算法信任研究提供新的理论基础；4.行为信任的深度解析，即模型突出机器行为对人机信任的影响，例如机器拒绝人类请求时对“感知被信任”的负面影响，揭示信任失调的情感与行为后果。综上，本模型以动态性、双向性和情感传递性为核心特征，全面展现了人机信任的复杂机制。

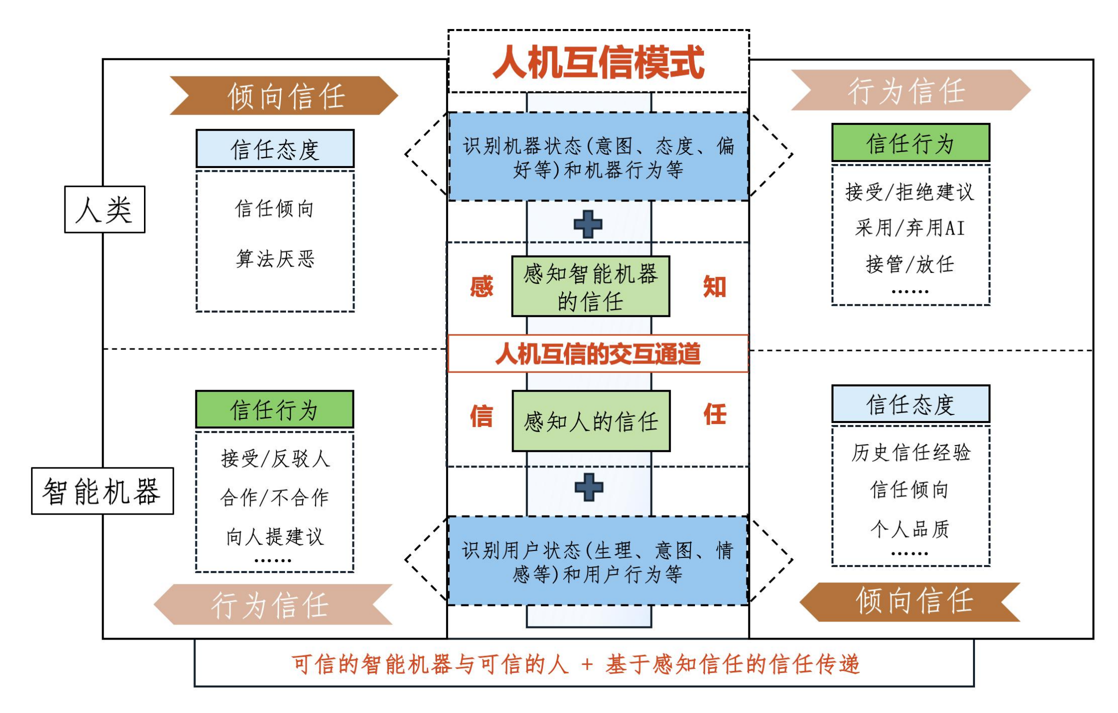


图 1 人机双向信任的理论模型

3.2 倾向信任

在人机信任研究中，倾向性信任指用户基于固有信任倾向和认知特质，在与技术或系统（如人工智能、自动化设备）实际交互前表现出的初步信任水平。其不依赖具体交互情境或经验，而由个性特征、总体态度、文化背景及先前经历等因素决定。Merritt 及其团队的研究(Merritt & Ilgen, 2008; Merritt, 2011; Merritt et al., 2013; Merritt et al., 2015)指出，倾向性信任主要受文化、年龄、性别和个性四个关键因素影响，具有普遍性和稳定性，是初始信任的基础。在我们的信任模型中，倾向性信任被视为构建人机信任的重要起点，为后续的感知信任和行为信任奠定基础，推动整个人机信任关系的建立与发展。

算法是人工智能发展的核心驱动力，其在智能机器中的重要性随着技术从自动化到人工智能转型而日益突出。以往人机信任研究关注用户对自动化设备或 AI 产品的信任，较少探讨对智能系统底层算法的态度。在人工智能时代，基于大数据和机器学习构建的算法展现了强大的能力，但也伴随不可预测性和不可解释性，增加了用户理解难度，并可能导致信任动摇，尤其面对决策不透明和公平性问题上(罗映宇 等，2023)。在学术界，用户拒绝接受或使用算法建议和服务的行为或行为倾向，被称为算法厌恶（Algorithm Aversion）(Dietvorst et al., 2015; Dietvorst et al., 2018; Mohlmann et al., 2021)。算法厌恶已在多个领域（如医疗、电子商务、自动驾驶、法律等）显现，并与个体因素如性别、年龄、大五人格等相关(罗映宇 等，2023)。研究表明，算法厌恶对人机信任产生负面影响(Castelo et al., 2019; Reich et al., 2023)，并显著影响用户采纳算法建议(Allen & Choudhury, 2022)及团队中的组织关系(陈龙, 2020; Mahmud et al., 2022)。尽管其影响已被证实，但在现有的信任模型中尚未充分考虑。鉴于算法在人工智能时代的重要性及其在信任过程中的作用，本研究将算法厌恶纳入人机双向信任理论模型，作为倾向性信任的组成部分。我们假设：算法厌恶倾向与信任倾向负相关，即算法厌恶个体的初始信任水平较低。同时，算法厌恶对感知信任和行为信任的影响机制可能与信任倾向类似，但效果相反。通过将算法厌恶纳入模型，我们期望能更全面地理解人机信任的形成机制，并揭示信任演化过程。但算法厌恶对感知信任与行为信任的具体作用机制及其长期影响仍需进一步实证研究，这将为算法设计和人机交互优化提供有价值的理论支持。

对于智能机器对人类的倾向性信任，齐玥等人(2024)提出，智能机器对人类的信任倾向在 AI 信任初期由信任经验和倾向决定，主要源于系统设计者对用户的信任假设。我们的模型扩展了这一观点，认为智能机器的信任倾向不仅受到设计者影响，还可能随着智能水平的提升逐渐呈现出类似人类的信任模式。具体来说，智能机器的初始信任可能反映设计者的倾向，但随着交互的增多，机器会根据用户行为和反馈调整信任倾向水平。例如，机器可能根据用户的行为模式和合作意愿修正对用户的信任。Johnson 和 Obradovich (2022)的实验发现，智能系统（如 ChatGPT）对人类群体具有一定的信任倾向，表明智能机器不仅继承设计者的信任设定，还可能自主调整信任。我们认为，智能机器对人类的信任倾向与人类的类似，主要由设计者的信任逻辑、机器学习经验和人类历史行为共同作用。智能机器通过学习用户行为和品质，逐步优化信任倾向，这为人机双向信任的动态演化提供了新的研究视角。

3.3 感知信任

感知信任是人机信任双向性的重要体现，也是人机双向信任的重要交互通道。在人际团队研究中，感知信任被视为团队合作与效能的关键因素之一。感知信任是指被信任者对信任者传递的信任大小的主观评估，它不仅受到信任者表现出的信任程度的影响，还依赖于被信任者对这种信任的接受和回应。在人类团队中，信任的传递主要通过双方对彼此信任的感知来实现(Alhaji et al., 2024)。被信任感是人类团队中广泛提及的一种信任形式(Ding & Liang, 2018; Rotter, 1967)。Baer 等人(2021)通过评估员工的信心以及他们对上司是否愿意接受其弱点的看法，将这些因素作为员工是否感到被信任的指标，发现员工的被信任感与上司的支持和接纳度密切相关。Gillespie(2012)则通过询问员工的上司是否愿意在工作中依赖他们，以及是否愿意分享个人观点和敏感信息，来衡量员工对上司信任的感知。Lau 和 Lam(2008)发现，相较于员工对上司的信任，上司对员工的信任是否被员工感知到，对员工的表现和态度有更强的影响。个体还会基于被信任的感知对彼此做出反应(Salamon & Robinson, 2008)。这些研究表明，信任感知在人类组织中具有良好的心理测量特性，并且是预测个体在团队环境中行为及合作意愿的重要指标(Baer et al., 2021; Lau et al., 2014; Simons et al., 2022)。

参考人际信任的传递理论，我们在人机双向信任模型中，将感知信任视为信任传递的重要通道。无论 AI 是否具备情感信任，它都会基于特定算法对人类做出判断并采取行动，传递类似于“信任信号”的感知信号。感知信任包括以下两方面主要内涵：1.人/智能机器对对方状态和行为的感知。例如，人类用户通过 AI 的表现、反应速度、决策透明度等因素感知其可靠性和有效性。同样，智能机器也会通过观察人类行为（如决策模式、反应速度、合作意图等），感知人类的可靠性和一致性。2.人/智能机器对对方信任的感知。例如，人类用户通过智能机器的决策方式、互动回应等，感知到机器是否对自己表现出信任。同样，智能机器也可能通过人类对其行为的反馈，感知到人类是否信任它。这种感知直接影响信任的形成，个体根据对方行为的感知调整信任水平。感知信任在信任的形成和维系中起到关键作用，是人机协作和提升信任质量的重要因素，尤其在复杂交互环境中为信任的动态构建提供了新的视角。

3.4 行为信任

行为信任指个体在特定情境下基于倾向信任和感知信任的综合加工，所做出的实际依赖或合作行为，是信任关系中的行动体现，也是所有信任模型的核心维度。高在峰等人(2021)和齐玥等人(2024)都强调了行为信任的重要性。高在峰等人将其与实时信任结合，认为情境和系统特征影响系统表现，从而形成实时信任；齐玥等人则定义为被信任方是否执行决策及其执行结果对系统的影响。行为信任体现在三个方面：1)依赖行为，指人类基于对机器的信任决定是否交任务。例如，在自动驾驶系统中，用户基于对系统的信任，决定是否将驾驶控制权交给汽车，以及对自动驾驶汽车的使用意愿(Zhang et al., 2019)。2)合作行为，指人类是否与智能机器协作。例如，在医疗、金融和客服等领域，用户依据对智能系统的信任选择参

与合作并共享敏感信息(Hancock et al., 2011)。3)建议采纳行为，指用户是否接受机器的建议。智能机器的信任行为表现出类似人类的模式，例如，用户采纳系统提供的产品推荐、诊断建议等(Xie et al., 2024)。

在智能机器对人类的信任行为方面，新型人机关系强调智能机器不仅是工具，还可以作为具备一定决策能力和智能化程度的主体。这种角色转变使智能机器逐渐表现出类似人类的信任行为模式。Xie 等人(In Press)证明，当机器通过判断拒绝人类的建议或指令时，会使用户产生未被机器信任的感受，从而降低用户对机器的使用意愿和态度。本文认为，智能机器对人类的信任行为主要包括依赖行为、合作行为以及建议采纳行为，并与人类对智能机器的信任行为呈现一定的对称性。然而，智能机器通过何种行为向人类传递信任信号仍是一个值得进一步研究和探索的方向。

综上，本文将人机互信划分为倾向信任、感知信任和行为信任三个阶段，展现了信任在交互中的动态演化特性，并结合各阶段特性提出适合的信任测量与计算方法，为人机信任领域的理论与实践发展提供系统框架支持。

4. 人机互信的测量和计算建模方法

提出理论模型的目的是基于不同阶段的测量特性，发展针对性的人机互信测量与计算建模方法，从而实现人机互信的动态监测与校准。已有大量研究探索了人类对智能机器信任度的测量(de Visser et al., 2018; Jian, 2000; Lee & See, 2004; Yu et al., 2018)，信任测量主要包括三种方式：主观量表报告、基于生理的测量和基于行为的测量。本章节将在总结现有测量方法的基础上，结合人际信任测量的经验，提出适用于人机双向信任的测量框架与方法。研究将关注以下几个方面：为倾向信任、感知信任和行为信任阶段开发阶段性测量工具；探索多维度、多层次的测量方法，融合主观报告、生理信号和行为数据，构建动态监测与校准系统；借鉴人际信任量化方法，设计适应人机交互特征的信任建模工具。最终，本研究旨在为人机双向信任的测量与建模提供系统化、可操作的理论与方法框架，为实现动态评估与智能化调节提供基础。

4.1 倾向信任的度量

根据第 3 章的理论模型框架，人对机器的倾向信任包括信任倾向和算法厌恶两个层面。信任倾向的测量普遍采用 Merritt 等人(2013)开发的信任机器倾向量表(Propensity to Trust Machines Scale)，该量表包含 6 个题目，广泛用于机器、技术及人工智能的信任倾向研究(Montag et al., 2023; Kohn et al., 2021)，为信任倾向量化提供了可靠的基础。算法厌恶的测量一般通过问卷调查参与者在特定情境下的选择倾向。例如，Reich 等人(2023)通过问题“你更信任谁来预测人格特质？（0=人类，100=算法）”直接衡量参与者在特定场景下对算法的厌恶。此外，Shariff 等人(2021)通过比较相同场景下人们对算法和人类操作者的期望差异来评估算法厌恶。随着算法厌恶现象的普遍性，学者们呼吁开展更系统的量化研究，包括开发算法厌恶倾向量表和构建算法厌恶程度指数(罗映宇 等，2023)。这些量化工具为信任测量与

建模的过渡提供了支持。我们在理论模型中整合了这些研究进展，作为信任倾向层面的核心组成部分。

相比之下，机器对人的信任倾向研究仍处于初步阶段。现有研究如 Johnson 和 Obradovich (2022)通过信任博弈观察 ChatGPT 的行为，尝试衡量智能机器对人类的信任倾向。信任博弈是经典的信任和合作测量方法，通过资金分配行为揭示参与者是否选择信任他人，及是否值得信任(Berg et al., 1995)。在信任博弈中，委托人(Trustor)决定是否信任另一方，受托人(Trustee)决定是否回报委托人的信任。委托人可以选择将部分或全部资金委托给受托人。委托的资金会被实验者以一个固定倍数(如 3 倍)增值，委托人转移的金额反映其信任水平，受托人返还的金额反映其信任回报或可信度水平。基于这一框架，Johnson 和 Obradovich 通过让 ChatGPT 作为委托人(Trustor)与人类互动，测量其对人类的信任倾向。他们发现，在适当激励条件下，ChatGPT 表现出对人类总体的信任倾向。虽然这种方法通过行为数据衡量信任倾向，但其局限性在于尚不能解释智能机器信任的成因、动机及影响因素。当前也缺乏类似于人类信任倾向的主观测量工具。

基于现有研究和测量方法，我们提出以下建议：参考现有信任倾向量表框架，设计适用于智能机器对人类信任倾向的量表；利用大语言模型的评估能力，通过询问大语言模型(如 ChatGPT)完成量表填写和测量。进一步，结合行为数据与模型输出开发新的测量方法。通过这些方法的探索和实践，能够进一步完善人机双向信任的测量与建模，为动态信任的监测和校准提供更加科学和有效的支持。

4.2 感知信任的度量

感知信任的度量是人机交互领域的重要研究方向，尤其是在探索人机双向信任的交互方面。主观量表作为一种广泛使用的信任测量手段，已经在多个领域得到了验证。Alsaid 等人(2023)回顾了信任测量工具，指出 Jian (2000)的 12 项信任量表广泛用于自动化系统信任评估。此外，Merritt 等人(2011)开发了多维度信任量表，将信任细化为可靠性和适用性等维度。Hoffman 等人(2013)则开发了动态信任量表，捕捉信任随时间或情境变化的特点。除通用量表外，研究者还针对特定应用场景开发了自我报告量表，涉及自动驾驶(Garcia et al., 2015；Robertson, 2021)、计算机(Madsen & Gregor, 2000)、机器人(Hancock et al., 2021)、自动化系统(Schaefer et al., 2016)等领域。这些量表具有以下共性：1. 任务后评估——关注对机器当前行为的感知，而非整体态度；2. 场景依赖性——量表题目与机器行为密切相关。

在机器对人类的信任研究中，机器缺乏人类情感，使得其信任机制尚不现实。但在组织管理领域，信任测量可以通过下属感知被信任的方式间接评估(Baer et al., 2021)。这一思路为机器对人类信任的测量提供了启发。本研究建议从人类感知被机器信任的角度出发，测量智能机器通过语言、行为传递的信任。具体步骤可以是：收集两类量表，一类用于衡量人类组织成员之间的信任，另一类用于评估人类对人工智能、机器等的信任。在收集原始题库后，应根据量表题目的信任测量视角（感知行为信任）进行筛选，剔除那些测量倾向性信任等整

体态度的题目。随后，研究者可以通过将量表题目从主动表述转换为被动表述，或交换主语和宾语，进行相应修改，以用于测量参与者的感知被信任程度。例如，可以设计题目："在此场景中，我认为智能机器信任我的决定。"参照上述路径，Xie 等人(In Press)开发并验证了一套用于测量人类感知被信任的问卷，评估自动驾驶汽车和 AI 在接受或拒绝人类建议时的信任感知。这项研究表明，积极的感知被信任显著促进人类对 AI 的信任及技术接受意愿。用户感知 AI 信任其判断时，更容易增强对 AI 的信任，提升技术接受意愿。Xie 等人的研究不仅完善了感知被信任的量化方法，也为优化人机交互设计提供了框架，推动了信任测量和技术接受的学术发展。

4.3 行为信任的度量

在人机信任研究中，基于行为的度量方法通过观察用户与系统的交互行为来评估信任水平，能够反映用户对系统能力、可靠性及意图的信任。常见的行为信任指标包括：系统采纳率，如在多轮交互中持续使用系统的频率(Chancey et al., 2017)；人机团队的绩效表现，即任务完成效率或成果质量(de Visser et al., 2016)；决策时间，即用户在决策过程中的犹豫或果断程度(Payre et al., 2017)；干预系统的次数，即用户在任务执行中对系统行为进行调整或干预的频率(Techer et al., 2019)；对系统的监控和输入，即用户在任务执行中对系统的监控强度和操作行为(Lee et al., 2021)等。另一类方法通过综合评估人类与机器交互过程中的一系列行为表现来衡量信任。例如，非语言行为，包括：肢体语言——如触摸脸部、双臂交叉、身体后倾等(Lee et al., 2013)；面部表情和眼神注视——如通过表情分析或视线追踪判断用户对系统的信任水平；空间距离和声音——如用户与系统的交互距离或语音语调的变化等(Ambady & Weisbuch, 2010)。在自动驾驶领域，综合多种行为指标可以全面评估驾驶员对自动驾驶汽车的信任。典型的行为信任指标包括：接管行为——用户在任务执行中是否主动接管系统(Pakdamanian et al., 2021)、车速和油门控制(Feng et al., 2016)，以及注视程度——驾驶员对系统界面的关注频率和时长(Strauch et al., 2019)等因素，均被用来衡量驾驶员在模拟驾驶或实际驾驶过程中的信任水平。这些指标能够反映驾驶员在面对自动驾驶系统时的依赖程度和心理反应。总的来说，在衡量人类对机器行为信任的过程中，研究者们普遍采用评估人类在不同情境下对机器响应行为的度量方法。

借鉴这些方法，机器对人类的信任可以通过分析不同人机协作场景下机器对人类行为的反应进行。例如，在与大语言模型（如 ChatGPT）协作时，观察机器如何根据人类的表现调整行为以评估机器的信任度。此外，还可以通过为机器设定一定的评价规则，来定义和衡量机器对人类的信任。例如，在 AI 简历筛选系统中，根据候选人表现评估系统的信任度；在个人信用评级中，分析历史还款行为量化 AI 对用户信用的信任；在驾驶评分系统中，根据驾驶员的安全或冒险行为评估机器的信任。Jorge 等人(2024)通过设计一个虚拟超市寻货任务的游戏，对 AI 信任进行了结构化测量。他们将参与者在单位时间内完成的寻货数量作为

能力指标，参与者与 AI 的合作意愿作为善意指标，以及参与者对 AI 撒谎的频率作为正直指标，综合这三个指标，以此作为 AI 对人类信任的评价标准。

4.4 综合动态双向信任的度量

以上的人机信任测量方法主要集中于主观心理或行为层面的单一维度，通常仅涉及人机双方中一方对另一方的信任。在实际使用中，如果需要测量双向信任，通常需要分别使用这些工具，测量双方的信任，再将结果合并以反映双向信任。这种方法虽然在一定程度上解决了双向信任的测量问题，但结果仍然缺乏统一性，且所需工具繁琐。为弥补这些不足，研究者提出了综合动态双向信任的度量方法，结合主观数据（如问卷反馈）、行为数据（如交互日志）和生理数据（如心率、皮肤电导等）等多模态信息，实现对信任关系的动态追踪与评估。这种多模态方法不仅能够揭示信任在交互过程中的变化，还能更全面地反映人机信任的复杂性与互动性。表 2 总结了部分典型的人机信任综合计算度量模型，我们旨在通过对这些模型的分析，提出一种基于统一数据集的预测双向信任动态变化的方法。该方法结合现有的信任测量技术，整合主观、行为及生理数据，进一步探讨如何通过这些数据有效预测和评估人机双向信任在交互过程中的演变。

表 2 典型人机信任的综合计算度量模型

研究	方法	建模数据	信任动态性	应用场景
Sadrfaridnour et al. (2016)	机器学习	初始信任、机器人表现、人类表现	动态模型	机器人人机协作
Bonneviot et al. (2021)	机器学习	行人行为、情绪水平	动态模型	行人与车辆
Li & Lee (2022)	效用模型	情景结构、战略行为、目标	动态模型	一般人机交互
Kamaraj et al., (2023)	机器学习	驾驶风格、速度、油门刹车等控制数据	静态模型	自动驾驶
Kamaraj et al., (2024)	机器学习	个人特质：驾驶风格、风险寻求；行为数据：油门、刹车控制等	静态模型	自动驾驶
Hu et al. (2024)	机器学习	驾驶员性格、历史经验、对系统感知	静态模型	自动驾驶
Li et al., (2024b)	效用模型	能力、信任信念	动态模型	一般人机交互
Li et al., (2024a)	机器学习	信任词汇、语音语调、响度	静态模型	语音对话
Yi et al. (2024)	机器学习	皮肤电、心电图、接管行为	动态模型	自动驾驶

在人对机器信任的综合计算模型中，研究者通常结合心理、生理、行为数据，并利用机器学习算法对人类对机器的信任进行量化和预测。这些模型的建模逻辑基于在人机实时交互过程中观察到的信任因素表征与对应的客观测量指标及时间动态特性之间的关系(Agreste et al., 2015; Gebru et al., 2022; Uggirala et al., 2004;)。自动驾驶领域的驾驶员信任监测是人机信任计算建模的重要研究方向之一，通常采集以下实时监控数据：手部动作（如方向盘操作频率和力度）、眼部动作（如凝视时间和视线轨迹）、非驾驶活动（如驾驶员分心行为）、系统使用（如自动驾驶功能的开启或关闭频率）、生理信号（如脑电、皮肤电和心率变化等）

(Avetisyan et al., 2023; Michelaraki et al., 2023; Qu et al., 2023)。与这些客观数据对应的信任判断指标具有实时性和动态性，常见的包括接管活动、关键事件的频域和时域特征、注视兴趣区切换等(Dong et al., 2010; Fernández et al., 2016; Lee et al., 2023; Qu et al., 2023)。Yu 等人(2021)通过结合客观和主观指标，提出了预测驾驶员对自动驾驶汽车信任的模型，并通过比较驾驶员手部位置和运动，评估信任水平。Avetisyan 等人(2024)将驾驶员的性格、初始信任和动态信任等因素纳入人车信任监控模型，揭示了车辆故障对信任的影响。Yi 等人(2024) 通过融合皮肤电、心电图等生理信号与驾驶员接管行为的交互体验，利用标签平滑的卷积神经网络（CNN）和长短期记忆网络（LSTM），建立了针对驾驶员对自动驾驶汽车信任的实时识别模型。这一模型能够动态捕捉驾驶员的信任状态，为信任的精准评估提供了新的技术手段。Zhang 等人(2024)通过脑电信号评估驾驶员对自动驾驶汽车信任的细微变化，提出了新的评估方法。

以上方法的共同特点在于，它们融合了心理、生理，尤其是行为的多模态数据，进行动态监测，并运用机器学习算法捕捉和预测人类对机器的信任水平。值得注意的是，尽管心理数据通过有针对性的量表进行测量，具有一定的特异性，但生理和行为数据通常涉及对个体行为的全面监控，信任的测量依赖于对通用数据的采集与定义，从而推导出信任预测的方法。基于这一思路，我们可以对人机交互过程中的通用数据进行分别定义，从而实现一套数据同时测量双向信任的构想。这种方法通过对双方信任的相互作用进行系统建模，使得单一数据集既能捕捉个体对机器的信任，也能反映机器对个体的信任。生理数据具有双向性，例如，在驾驶过程中，注视点信息和面部表情不仅可以用来推断驾驶员对车辆的不信任（如注视兴趣区从驾驶次任务切换到主驾驶任务，可能反映对车辆的不信任）(Fernández et al., 2016; Qu et al., 2023)。同时，这些数据还可以揭示驾驶员的疲劳、认知负荷、注意力等状态(Lu et al., 2016)。行为数据同样具有双向性，驾驶员的行为不仅反映了人对机器的信任，也能用于定义机器对人的信任。例如，特斯拉的驾驶员安全评分体系通过用户的急刹车、急转弯、危险跟车等五种手动驾驶行为，以及自动驾驶使用中的接管时间等指标，综合评估驾驶员的安全驾驶能力。这些指标可以被视为机器对驾驶员信任的体现。接管行为的反应时间也被研究者定义为衡量人对自动驾驶系统信任的指标。较长的接管时间可能表明驾驶员对系统的信任较高，较短的接管时间可能意味着不信任(Dong et al., 2010; Lee et al., 2023)。

基于以上方法，研究者可以通过对机器对人信任的理论模型进行结构化定义，设计统一的生理、行为和心理监测系统，以动态预估人机双向信任水平。这种综合性方法不仅能够精准捕捉人机交互中的信任动态，还能为人机系统的优化与适应性调整提供有力支持。未来，这种双向信任建模方法将进一步推动智能系统的可靠性与人机协作效率的提升，为智能系统的创新和发展提供坚实的基础。

5. 应用案例研究与未来研究方向

5.1 应用案例研究

人机互信的研究在多个关键领域中具有重要意义，能显著提升协作效率，并对事故预防、安全分析和主动干预产生积极作用。通过互信，机器能准确理解和支持人类决策，减少冲突和“人机互搏”现象(Prahl et al., 2022)。例如，在自动驾驶汽车与人类司机协同驾驶的场景中，当自动驾驶系统具备较高的驾驶能力，而人类驾驶员因能力受限或状态不佳无法胜任驾驶任务时，机器需要准确评估驾驶员的能力状态，以决定是否介入或接管控制(Ebnali et al., 2019; Lu et al., 2016)。在这一过程中，自动驾驶系统对驾驶员能力的评估体现了系统对人类的信任。此外，当系统通过提醒、主动干预或反馈驾驶能力和安全评分时，会直接引发驾驶员的信任感。这种感知不仅影响驾驶员对系统的信任，还对技术接受度和建议采纳意愿产生作用。因此，基于人机互信，特别是人类的信任感知，设计合理的系统提醒机制，能够平衡主动介入可能引发的不适与安全系统带来的收益。合理的人机信任设计有助于优化驾驶员与自动驾驶系统的交互体验，提升安全性(Seet et al., 2020)，增强用户对系统的接受度和依赖性，为更高效的协同驾驶提供支持。

在航空飞行领域，随着技术可靠性不断提升，飞行员在特殊情境下的应对能力对飞行安全至关重要。调查显示，75%以上的民航事故源于人为因素，其中 41%与非预期事件处置不当相关(Mathavara & Ramachandran, 2022)。当飞行员遭遇急性应激反应时，可能会出现一系列生理、心理和行为反应，如激素分泌增加、呼吸急促、心跳加快、情绪紧张、认知失调等，严重时甚至丧失对飞机状态的认知技能，从而导致灾难性后果(Walmsley & Gilbey, 2017; Wiggins et al., 2014)。在这种情况下，民航自动驾驶系统（APS）不仅能辅助飞行员长时间执行飞行任务，还能在飞行员状态不佳时提供关键安全保障，弥补其可能的错误或判断失误。通过人机双向信任理论模型和测量手段，可以实时监控飞行员的能力状态，分析其错误行为并生成综合评分（可视为系统对飞行员的初始信任）。这一评分机制不仅有助于系统在必要时及时介入，还能为后续培训提供依据。同时，基于双向信任模型的功能协调分配，有助于避免“人机互搏”现象，从而降低航空事故的发生率(He et al., 2023; Parrnell et al., 2021)。这种双向信任机制为飞行安全保障和人机协作优化提供了重要支持。

上述案例从不同领域展示了机器对人的信任对人类的深远影响。机器的信任不仅支持人类行为，还伴随着人类对机器信任水平、态度及使用意愿的动态变化。这进一步强调了构建可靠人机互信体系在优化协作效率和提升安全性方面的关键作用。

5.2 未来研究方向

基于上述人机双向信任的理论框架与测量方法分析，我们认为在人工智能时代人机互信的研究方面，以下 3 个方面的研究值得进一步关注。

（1）**机器对人类信任的测量工具开发。**在人机交互过程中，机器对人类信任的测量工具至关重要，尤其是在机器应用于特定任务和场景时。现有的测量工具需要通过问卷调查、行为实验等方式进行验证与完善。同时，工具在不同场景和任务环境下的适用性及其调整问

题也应受到关注。例如，智能助手、自动驾驶汽车和工业机器人等领域有不同的信任需求和评估标准。未来研究应致力于建立更加细化、精准的测量工具，以提升人机协同效果。

（2）**人类对机器信任和反馈干预的接受意愿。**人类对机器的技术接受度和使用意愿是人机团队研究中的重要议题(Wong et al., 2024)。尤其是当机器对人类进行评分、输出信任或对行为提出建议与干预时，能否接受这些反馈及其接受程度直接影响人机交互的质量和效果。当前，机器信任的接受意愿和态度尚未有系统研究，特别是不同类型反馈（如积极或消极反馈）如何影响人类接受度，及如何通过优化交互设计提高接受程度的问题，仍需进一步探索。

（3）**机器信任对人机协作绩效及心理、态度、行为的影响机制。**机器信任不仅影响人机协作结果，还通过潜移默化的方式影响用户的心理状态、态度和行为。尽管已有研究关注机器对人类的信任，但多停留在理论模型阶段，缺乏系统实证研究。研究应深入探讨机器信任对人类心理状态的影响（如自信心、自主性或被控制感等），以及机器信任是否会引发人类对机器长期态度的变化。此外，机器信任与人类信任的对齐问题也需研究，特别是机器是否存在“过度信任”或“信任不足”的现象，以及这些现象如何影响人机协作效果，均是未来的研究方向。

参考文献

陈龙. (2020). “数字控制”下的劳动秩序——外卖骑手的劳动控制研究. *社会学研究，35*(06), 113–135+244.

高在峰，李文敏，梁佳文，潘晗希，许 为，沈模卫. (2021). 自动驾驶车中的人机信任. *心理科学进展，29*(12), 2172–2183

罗映宇，朱国玮，钱无忌，吴月燕，黄 静，杨 智. (2023). 人工智能时代的算法厌恶：研究框架与未来展望. *管理世界，23*(10)，205–227

齐 玥，陈俊廷，秦邵天，杜 峰. (2024). 通用人工智能时代的人与 AI 信任. *心理科学进展，32*(12)，1–13

许为, 高在峰, 葛列众. (2024). 智能时代人因科学研究的新范式取向及重点. *心理学报，56*(3), 363–382.

许为, 葛列众. (2020). 智能时代的工程心理学. *心理科学进展，28*(9)，1409–1425.

Agreste, S., De Meo, P., Ferrara, E., Piccolo, S., & Provetti, A. (2015). Trust networks: Topology, dynamics, and measurements. *IEEE Internet Computing*, *19*(6), 26–35.

Alhaji, B., Büttner, S., Sanjay Kumar, S., & Prilla, M. (2024). Trust dynamics in human interaction with an industrial robot. *Behaviour & Information Technology*, *1–23*. https://doi.org/10.1080/0144929X.2024.2316284

Allen, R., & Choudhury, P. (2022). Algorithm-augmented work and domain experience: The countervailing forces of ability and aversion. *Organization Science, 33*(1), 149–169.

Alsaid, A., Li, M., Chiou, E. K., & Lee, J. D. (2023). Measuring trust: A text analysis approach to compare, contrast, and select trust questionnaires. *Frontiers in Psychology*, *14*, 1192020.

Ambady, N., & Weisbuch, M. (2010). Nonverbal behavior. In Fiske S. T., Gilbert D. T., Lindzey G. (Eds.), Handbook of social psychology (pp. 464–497). Hoboken, NJ: John Wiley & Sons.

Avetisyan, L., Ayoub, J., Yang, X. J., & Zhou, F. (2024). Building contextualized trust profiles in conditionally automated driving. *IEEE Transactions on Human-Machine Systems*, *54*(6), 658–667.

Azevedo-Sa, H., Yang, X. J., Robert, L. P., & Tilbury, D. M. (2021). A unified bi-directional model for natural and artificial trust in human–robot collaboration. *IEEE Robotics and Automation Letters*, *6*(3), 5913–5920.

Babashahi, L., Barbosa, C. E., Lima, Y., Lyra, A., Salazar, H., Argôlo, M., ... & Souza, J. M. D. (2024). AI in the workplace: A systematic review of skill transformation in the industry. *Administrative Sciences, 14*(6), 127.

Baer, M. D., Dhensa-Kahlon, R. K., Colquitt, J. A., Rodell, J. B., Outlaw, R., & Long, D. M. (2015). Uneasy lies the head that bears the trust: The effects of feeling trusted on emotional exhaustion. *Academy of Management Journal, 58*(6), 1637–1657.

Baer, M. D., Frank, E. L., Matta, F. K., Luciano, M. M., & Wellman, N. (2021). Undertrusted, overtrusted, or just right? The fairness of (in) congruence between trust wanted and trust received. *Academy of Management Journal, 64*(1), 180–206.

Basu, C., & Singhal, M. (2016, March). Trust dynamics in human autonomous vehicle interaction: a review of trust models. *In 2016 AAAI Spring Symposium Series - Technical Report* (pp. 85–91). Palo Alto, CA: AAAI Press.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142.

Bonneviot, F., Coeugnet, S., & Brangier, E. (2021). Pedestrians-automated vehicles interaction: Toward a specific trust model. In Black, N. L., Neumann, W. P., & Noy, I. (Eds.), *Proceedings of the 21st Congress of the International Ergonomics Association (IEA 2021)*(Vol. 221, pp. 568–574). Springer, Cham.

Caldwell, S., Sweetser, P., O'donnell, N., Knight, M. J., Aitchison, M., Gedeon, T., ... & Conroy, D. (2022). An agile new research framework for hybrid human-AI teaming: Trust, transparency, and transferability. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *12*(3), 1–36.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research, 56*(5), 809–825.

Chancey, E. T., Bliss, J. P., Yamani, Y., & Handley, H. A. (2017). Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human Factors, 59*(3), 333–345.

Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction*, *31*(10), 692–702.

Chung, H., Holder, T., Shah, J., & Yang, X. J. (2024). Developing a team classification scheme for human-agent teaming. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 68*(1), 1394–1399.

Cuzzolin, F., Morelli, A., Cirstea, B., & Sahakian, B. J. (2020). Knowing me, knowing you: theory of mind in AI. *Psychological Medicine*, *50*(7), 1057–1061.

de Visser, E. J., & Pak, R., Shaw, T. H. (2018). From 'automation' to 'autonomy': The importance of trust repair in human-machine interaction. *Ergonomics, 61*(10), 1409–1427.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155–1170.

Ding, Y., & Liang, Z. (2018). Structural optimization and measurement of Chinese employees' perception of being trusted. In Striełkowski, W., Black, J. M., Butterfield, S. A., Chang, C.-C., Cheng, J., Dumanig, F. P., Al-Mabuk, R., Urban, M., & Webb, S. (Eds.), *Proceedings of the 2018 2nd International Conference on Management, Education and Social Science (ICMESS 2018)* (pp. 1392–1395). Atlantis Press. *Advances in Social Science, Education and Humanities Research.*

Dong, Y., Hu, Z., Uchimura, K., & Murayama, N. (2010). Driver inattention monitoring system for intelligent vehicles: A review. *IEEE Transactions on Intelligent Transportation Systems*, *12*(2), 596–614.

Earle, T. C., & Siegrist, M. (2006). Morality information, performance information, and the distinction between trust and confidence 1. *Journal of Applied Social Psychology*, *36*(2), 383–416.

Ebnali, M., Hulme, K., Ebnali-Heidari, A., & Mazloumi, A. (2019). How does training effect users' attitudes and skills needed for highly automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *66*, 184–195.

Fang, Z., Wang, J., Liang, J., Yan, Y., Pi, D., Zhang, H., & Yin, G. (2023). Authority allocation strategy for shared steering control considering human-machine mutual trust level. *IEEE Transactions on Intelligent Vehicles, 9*(1), 2002–2015.

Feng, F., Bao, S., Sayer, J., & LeBlanc, D. (2016). Spectral power analysis of drivers' gas pedal control during steady-state car-following on freeways. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60*(1), 729–733.

Fernández, A., Usamentiaga, R., Carús, J. L., & Casado, R. (2016). Driver distraction using visual-based sensors and algorithms. *Sensors*, *16*(11), 1805.

Garcia, D., Kreutzer, C., Badillo-Urquiola, K., & Mouloua, M. (2015). Measuring trust of autonomous vehicles: a development and validation study. In Stephanidis, C. (Ed.), *HCI International 2015-Posters' Extended Abstracts* (Vol. 529, pp. 610–615). Springer, Cham. *Communications in Computer and Information Science.*

Gebru, B., Zeleke, L., Blankson, D., Nabil, M., Nateghi, S., Homaifar, A., & Tunstel, E. (2022). A review on human–machine trust evaluation: Human-centric and machine-centric perspectives. *IEEE Transactions on Human-Machine Systems*, *52*(5), 952–962.

Georganta, E., & Ulfert, A. S. (2024). Would you trust an AI team member? Team trust in human–AI teams. *Journal of Occupational and Organizational Psychology, 97*, 1212–1241.

Gillespie, N. (2012). Measuring trust in organizational contexts: An overview of survey-based measures. In Lyon, F., Möllering, G., & Saunders, M. (Eds.), *Handbook of research methods on trust* (pp. 175–188). Edward Elgar Publishing.

Gunning, D., Vorm, E., Wang, Y., & Turek, M. (2021). DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters, 2*, e61. https://doi.org/10.1002/ail2.61

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527.

He, X., Nie, X., Zhou, R., Yang, J., & Wu, R. (2023). The risk-taking behavioural intentions of pilots in adverse weather conditions: an application of the theory of planned behaviour. *Ergonomics*, *66*(8), 1043–1056.

Hieronymi, P. (2008). The reasons of trust. *Australasian Journal of Philosophy, 86*(2), 21–236.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434.

Hoffman, R. R., Johnson, M., Bradshaw, J. M., & Underbrink, A. (2013). Trust in automation. I*EEE Intelligent Systems, 28*(1), 84–88.

Hu, C., Huang, S., Zhou, Y., Ge, S., Yi, B., Zhang, X., & Wu, X. (2024). Dynamic and quantitative trust modeling and real-time estimation in human-machine co-driving process. *Transportation Research Part F: Traffic Psychology and Behaviour*, *106*, 306–327.

Inga, J., Ruess, M., Robens, J. H., Nelius, T., Rothfuß, S., Kille, S., ... & Kiesel, A. (2023). Human-machine symbiosis: A multivariate perspective for physically coupled human-machine systems. *International Journal of Human-Computer Studies*, *170*, 102926.

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons, 61*(4), 577–586.

Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*(1), 53–71.

Johnson, T., & Obradovich, N. (2022). Measuring an artificial intelligence agent's trust in humans using machine incentives. arXiv preprint arXiv:2212.13371. https://doi.org/10.48550/arXiv.2212.13371

Jorge, C. C., Jonker, C. M., & Tielman, M. L. (2024). How should an AI trust its human teammates? Exploring possible cues of artificial trust. *ACM Transactions on Interactive Intelligent Systems*, *14*(1), 1–26.

Jorge, C. C., Tielman, M. L., & Jonker, C. M. (2022a). Artificial trust as a tool in human-AI teams. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 1155–1157). IEEE. https://doi.org/10.1109/HRI53351.2022.9889652

Jorge, C. C., Tielman, M. L., & Jonker, C. M. (2022b). Assessing artificial trust in human-agent teams: a conceptual model. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents* (IVA '22) (Article 24, pp. 1–3). Association for Computing Machinery. https://doi.org/10.1145/3514197.3549696.

Kamaraj, A. V., Lee, J., Domeyer, J. E., Liu, S. Y., & Lee, J. D. (2024). Comparing subjective similarity of automated driving styles to objective distance-based similarity. *Human Factors*, *66*(5), 1545–1563.

Kamaraj, A. V., Lee, J., Parker, J. I., Domeyer, J. E., Liu, S. Y., & Lee, J. D. (2023). Bimodal trust: High and low trust in vehicle automation influence response to automation errors. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 67*(1), 1144–1149. https://doi.org/10.1177/21695067231196244

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human Factors*, *65*(2), 337–359.

Kaur, D., Uslu, S., Durresi, A., Mohler, G., & Carter, J. G. (2020). Trust-based human-machine collaboration mechanism for predicting crimes. In Barolli, L., Amato, F., Moscato, F., Enokido, T., & Takizawa, M. (Eds.),

Advanced information networking and applications. AINA 2020 (Vol. 1151). Springer, Cham. *Advances in Intelligent Systems and Computing*.

Khastgir, S., Birrell, S., Dhadyalla, G., & Jennings, P. (2017). Calibrating trust to increase the use of automated systems in a vehicle. In Stanton, N., Landry, S., Di Bucchianico, G., & Vallicelli, A. (Eds.), *Advances in human aspects of transportation* (Vol. 484). Springer, Cham. *Advances in Intelligent Systems and Computing*.

Kintz, J. R., Banerjee, N. T., Zhang, J. Y., Anderson, A. P., & Clark, T. K. (2023). Estimation of subjectively reported trust, mental workload, and situation awareness using unobtrusive measures. *Human Factors, 65*(6), 1142–1160.

Kobayashi, G., Quilici-Gonzalez, M. E., Broens, M. C., & Quilici-Gonzalez, J. A. (2016). The ethical impact of the internet of things in social relationships: Technological mediation and mutual trust. *IEEE Consumer Electronics Magazine, 5*(3), 85–89.

Kohn, S. C., De Visser, E. J., Wiese, E., Lee, Y. C., & Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology*, *12*, 604977. https://doi.org/10.3389/fpsyg.2021.604977

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology, 50*(1), 569–598.

Lau, D. C., & Lam, L. W. (2008). Effects of trusting and being trusted on team citizenship behaviours in chain stores. *Asian Journal of Social Psychology, 11*(2), 141–149.

Lau, D. C., Lam, L. W., & Wen, S. S. (2014). Examining the effects of feeling trusted by supervisors in the workplace: A self - evaluative perspective. *Journal of Organizational Behavior, 35*(1), 112–127.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.

Lee, J. D., Liu, S. Y., Domeyer, J., & DinparastDjadid, A. (2021). Assessing drivers' trust of automated vehicle driving styles with a two-part mixed model of intervention tendency and magnitude. *Human Factors*, *63*(2), 197–209.

Lee, J. J., Knox, B., & Breazeal, C. (2013). Modeling the dynamics of nonverbal behavior on interpersonal trust for human-robot interactions. In *Trust and autonomous systems: Papers from the 2013 AAAI Spring Symposium* (pp. 46–47). AAAI.

Li, M., & Lee, J. D. (2022). Modeling goal alignment in human-AI teaming: a dynamic game theory approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 66*(1), 1538–1542. https://doi.org/10.1177/1071181322661047.

Li, M., Erickson, I. M., Cross, E. V., & Lee, J. D. (2024a). It's not only what you say, but also how you say it: Machine learning approach to estimate trust from conversation. *Human Factors*, *66*(6), 1724–1741.

Li, M., Kamaraj, A. V., & Lee, J. D. (2024b). Modeling trust dimensions and dynamics in human-agent conversation: A trajectory epistemic network analysis approach. *International Journal of Human–Computer Interaction*, *40*(14), 3571–3582.

Lu, Z., Happee, R., Cabrall, C. D., Kyriakidis, M., & De Winter, J. C. (2016). Human factors of transitions in automated driving: A general framework and literature survey. *Transportation Research Part F: Traffic Psychology and Behaviour*, *43*, 183–198.

Lyons, J. B., Wynne, K. T., Mahoney, S., & Roebke, M. A. (2019). Trust and human-machine teaming: A qualitative study. In Lawless, W., Mittu, R., Sofge, D., Moskowitz, I. S., & Russell, S. (Eds.), *Artificial intelligence for the Internet of everything* (pp. 101–116). Academic Press.

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In *Proceedings of the 11th Australasian Conference on Information Systems* (Vol. 53, pp. 6–8).

Mahmud, H., Islam, A. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change, 175*, 121390.

Mathavara, K., & Ramachandran, G. (2022). Role of human factors in preventing aviation accidents: An insight. In *Aeronautics-New Advances*(pp. 1–26). IntechOpen. https://doi.org/10.5772/intechopen.106899

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734.

McNeese, N. J., Demir, M., Cooke, N. J., & Myers, C. (2018). Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Human Factors, 60*(2), 262–273.

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors, 53*(4), 356–370.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors, 50*(2), 194–210.

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors, 55*(3), 520–534.

Merritt, S. M., Lee, D., Unnerstall, J. L., & Huber, K. (2015). Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human Factors, 57*(1), 34–47.

Michelaraki, E., Katrakazas, C., Kaiser, S., Brijs, T., & Yannis, G. (2023). Real-time monitoring of driver distraction: State-of-the-art and future insights. *Accident Analysis & Prevention*, *192*, 107241.

Möhlmann, M., Zalmanson, L., Henfridsson, O., & Gregory, R. W. (2021). Algorithmic management of work on online labor platforms: When matching meets control. *MIS Quarterly, 45*(4).

Montag, C., Kraus, J., Baumann, M., & Rozgonjuk, D. (2023). The propensity to trust in (automated) technology mediates the links between technology self-efficacy and fear and acceptance of artificial intelligence. *Computers in Human Behavior Reports*, *11*, 100315.

Mueller, F. F., Lopes, P., Strohmeier, P., Ju, W., Seim, C., Weigel, M., ... & Maes, P. (2020). Next steps for human-computer integration. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)* (pp. 1–15). Association for Computing Machinery.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies, 27*(5-6), 527–539.

Murphy, R. R. (2024). What will robots think of us? *Science Robotics, 9*(86), eadn6096. https://doi.org/10.1126/scirobotics.adn6096

Murphy, R., & Woods, D. D. (2009). Beyond Asimov: The three laws of responsible robotics. *IEEE Intelligent Systems, 24*(4), 14–20.

Nies, H. (2009). Key elements in effective partnership working. In Glasby, J., & Dickinson, H. (Eds.)*, International perspectives on health and social care: Partnership working in action (pp. 56–67).* Wiley-Blackwell.

Olson, D. M., & Xu, Y. (2021). Building Trust Over Time in Human-Agent Relationships. In *Proceedings of the 9th International Conference on Human-Agent Interaction* (pp. 193–201). Association for Computing Machinery. https://doi.org/10.1145/3472307.3484178

Pakdamanian, E., Sheng, S., Baee, S., Heo, S., Kraus, S., & Feng, L. (2021). Deeptake: Prediction of driver takeover behavior using multimodal data. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)* (Article 103, pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445563

Parnell, K. J., Wynne, R. A., Griffin, T. G., Plant, K. L., & Stanton, N. A. (2021). Generating design requirements for flight deck applications: Applying the perceptual cycle model to engine failures on take-off. *International Journal of Human–Computer Interaction, 37*(7), 611–629.

Payre, W., Cestac, J., Dang, N. T., Vienne, F., & Delhomme, P. (2017). Impact of training and in-vehicle task performance on manual control recovery in an automated car. *Transportation Research Part F: Traffic Psychology and Behaviour*, *46*, 216–227.

Pitardi, V., & Marriott, H. R. (2021). Alexa, she's not human but… Unveiling the drivers of consumers' trust in voice ‐ based artificial intelligence. *Psychology & Marketing*, *38*(4), 626–642.

Prahl, A., Leung, R. K. H., & Chua, A. N. S. (2022). Fight for flight: The narratives of human versus machine following two aviation tragedies. *Human-Machine Communication*, *4*, 27–42.

Qu, Y., Hu, H., Liu, J., Zhang, Z., Li, Y., & Ge, X. (2023). Driver state monitoring technology for conditionally automated vehicles: Review and future prospects. *IEEE Transactions on Instrumentation and Measurement, 72,* Article 3000920, 1–20.

Regli, C., & Annighoefer, B. (2022). An anthropomorphic approach to establish an additional layer of trustworthiness of an AI pilot. In *Software Engineering 2022 Workshops* (pp. 160–180). Gesellschaft für Informatik e.V. https://doi.org/10.18420/se2022-ws-17

Reich, T., Kaju, A., & Maglio, S. J. (2023). How to overcome algorithm aversion: Learning from mistakes. *Journal of Consumer Psychology, 33*(2), 285–302.

Robertson, I. W. T. (2021). The development and initial validation of the trust in self-driving vehicles scale (tsdv) (Doctoral dissertation). Rice University.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality, 35*(4), 651–665.

Sadrfaridpour, B., Saeidi, H., Burke, J., Madathil, K., & Wang, Y. (2016). Modeling and control of trust in human-robot collaborative manufacturing. In Mittu, R., Sofge, D., Wagner, A., & Lawless, W. (Eds.), Robust intelligence and trust in autonomous systems (pp. 115–141). Springer.

Salamon, S. D., & Robinson, S. L. (2008). Trust that binds: the impact of collective felt trust on organizational performance. *Journal of Applied Psychology, 93*(3), 593.

Schaefer, K. E., Billings, D. R., Szalma, J. L., Adams, J. K., Sanders, T. L., Chen, J. Y., & Hancock, P. A. (2014). A meta-analysis of factors influencing the development of trust in automation*, Implications for Human-robot Interaction*. Aberdeen: Army Research Laboratory.

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors, 58*(3), 377–400.

Seet, M., Harvy, J., Bose, R., Dragomir, A., Bezerianos, A., & Thakor, N. (2020). Differential impact of autonomous vehicle malfunctions on human trust. *IEEE Transactions on Intelligent Transportation Systems, 23*(1), 548–557.

Shariff, A., Bonnefon, J. F., & Rahwan, I. (2021). How safe is safe enough? Psychological mechanisms underlying extreme safety demands for self-driving cars. *Transportation Research Part C: Emerging Technologies, 126*, 103069.

Shi, Z., O'Connell, A., Li, Z., Liu, S., Ayissi, J., Hoffman, G., ... & Matarić, M. J. (2024). Build your own robot friend: An open-source learning module for accessible and engaging AI education. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, and Fourteenth Symposium on Educational Advances in Artificial Intelligence* (AAAI'24/IAAI'24/EAAI'24) (Article 2636, pp. 1–9). AAAI Press. https://doi.org/10.1609/aaai.v38i21.30359

Simons, T., Leroy, H., & Nishii, L. (2022). Revisiting behavioral integrity: Progress and new directions after 20 years. *Annual Review of Organizational Psychology and Organizational Behavior, 9*(1), 365–389.

Simpson, J. A. (2007). Psychological foundations of trust. *Current Directions in Psychological Science*, *16*(5), 264–268.

Strauch, C., Mühl, K., Patro, K., Grabmaier, C., Reithinger, S., Baumann, M., & Huckauf, A. (2019). Real autonomous driving from a passenger's perspective: Two experimental investigations using gaze behaviour and trust ratings in field and simulator. *Transportation Research Part F: Traffic Psychology and Behavior, 66*, 15–28.

Sycara, K., & Lewis, M. (2004). Integrating intelligent agents into human teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 46*, 413–417

Techer, F., Ojeda, L., Barat, D., Marteau, J. Y., Rampillon, F., Feron, S., & Dogan, E. (2019). Anger and highly automated driving in urban areas: The role of time pressure. *Transportation Research Part F: Traffic Psychology and Behaviour*, *64*, 353–360.

Uggirala, A., Gramopadhye, A. K., Melloy, B. J., & Toler, J. E. (2004). Measurement of trust in complex and dynamic systems using a quantitative approach. *International Journal of Industrial Ergonomics*, *34*(3), 175–186.

Ulfert, A. S., Georganta, E., Centeio Jorge, C., Mehrotra, S., & Tielman, M. (2024). Shaping a multidisciplinary understanding of team trust in human-AI teams: a theoretical framework. *European Journal of Work and Organizational Psychology*, *33*(2), 158–171.

Walliser, J. C., de Visser, E. J., Wiese, E., & Shaw, T. H. (2019). Team structure and team building improve human–machine teaming with autonomous agents. *Journal of Cognitive Engineering and Decision Making*, *13*(4), 258–278.

Walmsley, S., & Gilbey, A. (2017). Debiasing visual pilots' weather-related decision making. *Applied Ergonomics, 65*, 200–208.

Wang, J., & Moulden, A. (2021). AI Trust Score: A user-centered approach to building, designing, and measuring the success of intelligent workplace features. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 54, pp. 1–7). Association for Computing Machinery. https://doi.org/10.1145/3411763.3443452

Wang, Y., Wang, X., Tang, J., Zuo, W., & Cai, G. (2015). Modeling status theory in trust prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 29, 9460. https://doi.org/10.1609/aaai.v29i1.9460

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52*, 113–117.

Wiggins, M. W., Azar, D., Hawken, J., Loveday, T., & Newman, D. (2014). Cue-utilisation typologies and pilots' pre-flight and in-flight weather decision-making. *Safety Science, 65*, 118–124.

Wong, J. H., Chiou, E. K., Gutzwiller, R. S., Cook, M. B., & Fallon, C. K. (2024). Human-artificial intelligence teaming for the U.S. Navy: Developing a holistic research roadmap. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 68*(1), 380-385. https://doi.org/10.1177/10711813241260352

Xie, Y., Liu, Y., Zhou, R., Zhi, X., & Chan, A. H. (2024). Wait or Pass? Promoting intersection's cooperation via identifying vehicle's social behavior. *Accident Analysis & Prevention, 206*, 107724.

Xie, Y., Zhou, R., Chan, A. H. S., Jin, M., & Qu, M. (2023). Motivation to interaction media: The impact of automation trust and self-determination theory on intention to use the new interaction technology in autonomous vehicles. *Frontiers in Psychology, 14*, 1078438.

Xie, Y., Zhou, R., Chan, A.H.S., (In Press) Do You Trust Me? Measuring People's Perception of Being Trusted by AI in a Human–Agent Team. *International Journal of Human-Computer Interaction*

Yang, C., Zhu, Y., & Chen, Y. (2021). A review of human–machine cooperation in the robotics domain. *IEEE Transactions on Human-Machine Systems, 52*(1), 12-25.

Yi, B., Cao, H., Song, X., Wang, J., Zhao, S., Guo, W., & Cao, D. (2024). How can the trust-change direction be measured and identified during takeover transitions in conditionally automated driving? Using physiological responses and takeover-related factors. *Human Factors*, *66*(4), 1276–1301.

Yu, B., Bao, S., Zhang, Y., Sullivan, J., & Flannagan, M. (2021). Measurement and prediction of driver trust in automated vehicle technologies: An application of hand position transition probability matrix. *Transportation Research Part C: Emerging Technologies*, *124*, 102957.

Yu, K., Berkovsky, S., Conway, D., Taib, R., Zhou, J., & Chen, F. (2018). Do I trust a machine? Differences in user trust based on system performance. In J. Zhou & F. Chen (Eds.), *Human and machine learning* (pp. 161–172). Springer.

Yuan, L., Gao, X., Zheng, Z., Edmonds, M., Wu, Y. N., Rossano, F., ... & Zhu, S. C. (2022). In situ bidirectional human-robot value alignment. *Science Robotics, 7*(68), eabm4183. https://doi.org/10.1126/scirobotics.abm4183.

Zhang, T., Tao, D., Qu, X., Zhang, X., Lin, R., & Zhang, W. (2019). The roles of initial trust and perceived risk in public's acceptance of automated vehicles. *Transportation Research Part C: Emerging Technologies, 98*, 207–220.

Zhang, T., Yang, J., Chen, M., Li, Z., Zang, J., & Qu, X. (2024). EEG-based assessment of driver trust in automated vehicles. *Expert Systems with Applications*, *246*, 123196.

Zhou, L., Paul, S., Demirkan, H., Yuan, L., Spohrer, J., Zhou, M., & Basu, J. (2021). Intelligence augmentation: Towards building human-machine symbiotic relationship. *AIS Transactions on Human-Computer Interaction, 13*(2), 243–264.

Zhang, T., Tao, D., Qu, X., Zhang, X., Lin, R., & Zhang, W. (2019). The roles of initial trust and perceived risk in public's acceptance of automated vehicles. *Transportation Research Part C: Emerging Technologies, 98*, 207–220.

The bidirectional trust in the context of new human-machine relationships

XIE Yubin[1,2], ZHOU Ronggang[1,3,4]

([1] *School of Economics and Management, Beihang University, Beijing* 100191, *China*)

([2] *Department of Systems Engineering, City University of Hong Kong, Hong Kong* 999077, *China*)

([3] *Key Laboratory of Data Intelligence and Management, Beihang University, Beijing* 100191, *China*)

([4] *Laboratory for Low-carbon Intelligent Governance, Beihang University, Beijing* 100191, *China*)

Abstract: With the rapid advancement of artificial intelligence (AI) technology, the frequency and complexity of human-machine interactions have significantly increased. Traditional human-machine trust models primarily focus on unidirectional trust, specifically human trust in machines. However, as intelligent systems gradually acquire autonomy and decision-making capabilities, the bidirectional nature of human-machine trust has emerged as a central topic of research. Building on a review of recent theoretical models of human-machine trust, this study proposes a bidirectional trust framework based on dispositional trust, perceived trust, and behavioral trust, with particular emphasis on the critical role of perceived trust as an interactive channel for mutual trust between humans and machines. Additionally, this paper systematically examines the latest advancements in trust measurement and computational modeling, with a specific focus on methods for assessing machine trust in humans and their practical implications. Finally, it identifies future research directions, aiming to provide new perspectives and a guiding framework for the theoretical development and technological application of human-machine collaboration.

Keywords: artificial intelligence, human-machine mutual trust, trust, trust measurement, human-machine teams